The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0737-8831.htm

INSTITUTIONAL REPOSITORIES IN CANADA

The CARL metadata harvester and search service

Mark Jordan
WAC Bennett Library, Simon Fraser University, Burnaby, Canada

Received October 2005
Accepted February 2006

Abstract

Purpose – To explain the background, functionality, and content of the CARL metadata harvester and search service, http://carl-abrc-oai.lib.sfu.ca/, and to outline plans for improving the service.

Design/methodology/approach – This case study employs simple statistical analyses to a set of harvested metadata.

Findings – This paper documents the use of unqualified Dublin Core (uDC) elements in the metadata harvested from the repositories participating in the CARL harvester, and identifies patterns in the use of that metadata. It also compares these findings with a similar study, and identifies areas for further research.

Research limitations/implications – This paper is limited to discussion of the characteristics of a relatively small set of metadata collected using the Open Archives Initiative Protocol for Metadata Harvesting. However, analyses reveal some patterns in the use of this metadata that are valuable in the development of best practices for repository implementers.

Practical implications – This paper documents the use of uDC elements by a specific community. Its findings will form a basis for developing mechanisms for improving the effectiveness of the metadata generated by that community and therefore the services built around that metadata.

Originality/value – While there are several other studies that take an approach similar to that taken in this paper, no one has yet studied this specific data set. More generally, this paper contributes a valuable case study to research on the implementation of the Open Archives Initiative Protocol for Metadata Harvesting.

Keywords Canada, Research libraries, Data handling, Digital libraries

Paper type Case study

Introduction

The Canadian Association of Research Libraries/Association des bibliothèques de recherche du Canada's Institutional Repository Metadata Harvester[1] collects metadata from nine repositories hosted at Canadian universities and provides a search interface to the aggregated metadata. Seven of the nine repositories are general institutional repositories (IRs), and the remaining two are dedicated to electronic theses and dissertations (ETDs). Not all repositories are at CARL members; any collection of scholarly information is eligible for inclusion and as new collections become available they will be added to the harvester.

This paper describes the harvester, analyzes the metadata being contributed to it, and explains the steps CARL is taking to improve the usefulness of this service.

Scope and history of the harvester

The harvester, which is hosted at Simon Fraser University Library, was originally established in March 2004 to enhance access to CARL members' IRs, and was

Library Hi Tech
Vol. 24 No. 2, 2006
pp. 197-210

0737-8831
DOI 10.1108/07378830610669574

established as an integral component of CARL's proactive support of IRs as described elsewhere in this issue.

Once per day, the harvester makes a request to each of the nine repositories using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[2], returns all new or updated records, and incorporates them into its database. This updating is fully automated and the new records are available immediately for searching after they are retrieved.

The repositories included in the harvester at time of writing (June 2005) are shown in Table I.

As can be seen from the Table I, DSpace is currently the most popular repository platform but others are also being used. Because implementation of the OAI-PMH is fairly mature in most repository platforms that incorporate it, new repositories can be added to the harvester easily regardless of which platform they use.

The Public Knowledge Project's (PKP) Open Archives Harvester Software[3] provides the basis for the CARL service. Simon Fraser University Library staff modified the PKP software slightly to remove some features that optimized it for use with the other PKP applications, the Open Journal Systems (OJS)[4] and the Open Conference Systems (OCS)[5]. These modifications will be described in detail later, but to summarize, the most important modification was to simplify the PKP software so that it was more consistent with the unqualified Dublin Core (uDC) metadata schema that is used by all of the participating repositories.

Growth of the harvester

Additions of new records to the CARL harvester are not currently logged, but a report is e-mailed to the harvester administrator each day indicating how many records have been added. Small numbers of new records (usually fewer than 10) are added almost every day during the automated harvesting cycle. Occasionally, larger numbers (sometimes hundreds) of new records are added in one day as repositories load large batches of documents. To put the growth of the harvester into perspective, at the end of September 2004, the harvester contained 3,243 records; in June 2005, it contained over 5,400.

Name	URL	Repository platform
Archimede Université Laval	http://archimede.bibl.ulaval.ca	Archimede[a]
Collection mémoires et thèses de l'Université Laval	www.theses.ulaval.ca	
DSpace@UCalgary.ca	https://dspace.ucalgary.ca	DSpace[b]
eCommons::Research (University of Winnipeg)	http://ecommons.uwinnipeg.ca	Eprints[c]
Mspace (University of Manitoba)	http://mspace.lib.umanitoba.ca	DSpace
Papyrus – Dépôt institutionnel numérique (Université de Montréal)	http://papyrus.bib.umontreal.ca	DSpace
Simon Fraser University Institutional Repository	http://ir.lib.sfu.ca	DSpace
T-Space (University of Toronto)	https://tspace.library.utoronto.ca	DSpace
University of Saskatchewan Electronic Theses and Dissertations	http://library.usask.ca/etd/	ETD-db[d]

Table I.
Repositories included in the CARL harvester

Sources: [a]www.bibl.ulaval.ca/archimede/, [b]www.dspace.org, [c]www.eprints.org, [d]http://scholar.lib.vt.edu/ETD-db/

The metadata

Analyzing the data collected by the harvester will allow us to:

- determine the frequency with which specific uDC elements are present in contributing repositories' records;
- determine how the elements are used across different repositories; and
- determine aspects of the aggregated metadata that may have negative effects on users' ability to retrieve records.

General models for this type of analysis include Jewel Ward's "Unqualified Dublin Core Usage in OAI-PMH Data Providers" (Ward, 2004) and Moen and Benardino's "Assessing Metadata Utilization: An Analysis of MARC Content Designation Use" (Moen and Benardino, 2003). Using information derived from this type of analysis, we can develop guidelines that will allow repositories to optimize the metadata they contribute to the CARL harvester, thereby improving users' ability to perform successful queries through the harvester and increasing the probability that users will be lead back to the local repositories and to the local repositories' potentially richer metadata and of course to the full text documents.

In order to perform simple statistical analyses on the uDC records contributed to the CARL harvester, the author wrote a script incorporating Ed Summer's Net::OAI::Harvester Perl module[6]. The resulting records were then loaded into a MySQL database so they could be counted in various ways. This approach simplifies statistical analyses because the simple harvester and the accompanying database are optimized for counting records and elements, whereas the same version of the metadata as it exists in the CARL harvester's database is optimized for searching.

Overall element frequency

As harvested on June 19, 2005, 5,445 records were available via OAI-PMH from all of the archives that contribute to the harvester. Table II lists the number of records retrieved from each repository.

The records harvested from the nine contributing repositories contain a total of 95,789 uDC elements. The frequency of each element, described as a percentage of all the elements in the total set of records, is given in Table III.

Five elements each account for more than 10 percent of all elements in the harvested metadata: date (14.6 percent of all elements), subject (13.7 percent), format (13.4 percent),

Name	Number of records	Percentage of total
T-Space (University of Toronto)	3,471	63.7
University of Saskatchewan Electronic Theses and Dissertations	780	14.3
DSpace@UCalgary.ca	483	8.9
Archimede Université Laval	257	4.7
Collection mémoires et thèses de l'Université Laval	179	3.3
Simon Fraser University Institutional Repository	129	2.4
eCommons::Research (University of Winnipeg)	99	1.8
Papyrus - Dépôt institutionnel numérique (Université de Montréal)	24	0.4
Mspace (University of Manitoba)	23	0.4

Table II.
Number of records in each repository

Element	Frequency	Percentage of total
Date	14,014	14.6
Subject	13,104	13.7
Format	12,799	13.4
Identifier	12,087	12.6
Creator	11,769	12.3
Description	5,880	6.1
Title	5,475	5.7
Type	5,419	5.7
Language	5,234	5.5
Publisher	4,286	4.5
Contributor	2,350	2.4
Rights	1,995	2.1
Source	787	0.8
Relation	590	0.6
Coverage	0	0

Table III.
Frequency of uDC elements in the harvested metadata

identifier (12.6 percent), and creator (12.3 percent). At the other end of the spectrum, two elements each account for less than 1 percent of all elements: source (0.8 percent) and relation (0.6 percent). There were no coverage elements at all in the harvested metadata.

Elements per record

Another way of looking at the harvested records is to determine the average number of each uDC element per record. The first column in the Table IV indicates the average number of times each element was used per record across all repositories, and the second column indicates how many repositories fell below that average. In other words, the second column indicates the number of contributing repositories whose records used an element less frequently than the average number of times that element is used in all records contributed to the CARL harvester. The third column indicates how many contributing repositories' records do not contain an element at all (Table IV).

Element	Average elements per record for all providers	Number of providers below average	Number of providers who do not include the element at all
Title	1	0	0
Creator	2.2	6	1
Subject	2.4	3	1
Description	1	1	0
Publisher	0.8	4	1
Contributor	0.4	2	3
Date	2.6	4	0
Type	1	1	0
Format	2.4	6	1
Identifier	2.2	7	0
Source	0.1	3	5
Language	1	0	1
Relation	0.1	1	5
Coverage	N/A	N/A	9
Rights	0.4	0	6

Table IV.
Average number of elements per harvested record

The single repository listed as not including any "creator" elements in its metadata has not yet upgraded from DSpace 1.1 to 1.2 (which does not map its internal "author" field to the Dublin Core "creator" element, as will be explained in detail below). However, that repository does actually include author elements in the native metadata schema of its IR, which map to creator elements in DSpace 1.2 and later.

No repositories include any "coverage" elements in their metadata; five of nine include no "source" element, five of nine include no "relation" element, and six of nine include no "rights" elements. All repositories include "title" "description" "date" "type" and "identifier" elements in at least some of their records.

Analysis of the data

The overall element frequency as detailed in Table III is not consistent with Jewel Ward's findings that creator (at 21.5 percent of all elements) was the most commonly used element in the 910,919 records she harvested records from 82 data providers. Identifier was the next most common (17.2 percent of all elements), followed by title (11.4 percent), date (11.1 percent), and type (10.7 percent). The remaining elements each comprised less than 6.6 percent of all harvested elements (Ward, 2004, p. 44). In the metadata collected from repositories contributing to the CARL harvester, date was the most common (14.6 percent of all elements), followed by subject (13.7 percent), format (13.4 percent), identifier (12.6 percent), and creator (12.3 percent), and the remaining elements comprised no more than 6.1 percent of all harvested elements.

It is not surprising that the element frequencies in the two sets of repositories are dissimilar since Ward used a considerably larger data set containing a larger variety of types of repositories, including a large number of disciplinary archives. It is interesting, though, that in both Ward's and the current data set, the relation, contributor, and source elements each comprised less than 1 percent of all elements.

It may be more useful to compare our findings with Ward's analysis of the number of data providers that never used particular elements (Ward, 2004, p. 44) (Table V).

Element	Ward's findings (out of 82 providers)	Current study (out of 9 providers)
Title	1 (1.2)	0 (0)
Creator	4 (4.9)	1 (11.1)
Subject	14 (17.1)	1 (11.1)
Description	23 (28)	0 (0)
Publisher	41 (50)	1 (11.1)
Contributor	50 (61.0)	3 (33.3)
Date	6 (7.3)	0 (0)
Type	10 (12.2)	0 (0)
Format	43 (52.4)	1 (11.1)
Identifier	7 (8.5)	0 (0)
Source	52 (63.4)	5 (55.5)
Language	39 (47.6)	1 (11.1)
Relation	66 (80.5)	5 (55.5)
Coverage	66 (80.5)	9 (100)
Rights	46 (56.1)	6 (66.6)

Notes: Figures given in parentheses are percentages

Table V.
Number of data providers who never included elements in their metadata

Again, Ward's data set is nearly ten times larger than ours, but in both Ward's study and in the current one, relatively few providers did not include title, creator, subject, date, type, and identifier. Conversely, both analyses found that high proportions of providers did not include source, relation, coverage, and rights. Wider gaps exist between the proportion of providers in Ward's study and in the current one with regard to format and language.

Sample element values from the CARL metadata

Within the elements that are common to all CARL repositories, there is considerable variation in how the elements are being used. For our analysis, we inspected the values of the date, type, description, subject, publisher, identifier, and relation elements as they appear in the harvested metadata. Two fields that are common to all repositories, such as creator and title, are used fairly consistently, which is not surprising since it is common practice in institutional and ETD repositories to populate these elements with information transcribed directly from the documents themselves.

Before beginning a discussing the variation in the contents of these elements, it is useful to remember that the metadata being analyzed is uD (or as it is frequently abbreviated, uDC). Support for this metadata schema is mandatory in all valid OAI repositories and harvesters, but the OAI-PMH standard does allow for and actually encourages harvesting metadata in other formats. All of the repositories submitting records to the CARL harvester are using uDC.

The use of uDC is a double-edged sword. The advantage of using this schema is that because software packages such as DSpace and Eprints support uDC by default, it is therefore easy to expose a repository's metadata via the OAI-PMH. Since, the harvester software is guaranteed to understand uDC, it is fairly easy to have harvesters retrieve valid records from a repository. The disadvantage of using uDC is that it is by design extremely simple, so simple in fact that mapping from a richer, more sophisticated record structure to uDC invariably leads to loss of information about the element contents. For example, a repository's native metadata schema might distinguish between "date published" and "date added to the repository" but because uDC is not able to distinguish between these two, they will both be represented simply as "date" in harvested metadata. Qualified Dublin Core, on the other hand, is capable of preserving much or in some cases all of the richness of native metadata schemas, but as explained above, most OAI-PMH repositories and harvesters use only uDC metadata since it is the default.

Date

Values of the "date" element fall into several patterns, most notably dates in the format 2005-01-13T16:26:15Z and dates in the format 2004-05-17 (the "Z" stands for "Zulu" or Greenwich Mean Time). The former is typical of timestamps and other dates assigned by the repository software itself. Many repository platforms assign this type of date when a document is ingested or submitted into the repository. The second pattern occurs at three levels of specificity, expressing either a year, month and a day (2004-05-17), just a year and a month (2004-2005), or only a year (2004).

It is likely that these dates, especially the ones that specify a year and month or just a year, are not generated automatically by the repository software but are publication or copyright dates that have been entered into the repository software's metadata creation tools by a faculty member, research assistant, or library staff member. Based solely on the value of the uDC element, we cannot tell what type of date we are dealing with since any

information about the meaning of the date has been stripped from the native metadata value as it was mapped to uDC for harvesting. All harvested date elements contain sufficient information for the CARL harvester to add them to its database properly, and the harvester does some automatic normalization of dates as it adds them to the database (it truncates all date elements to the yyyy-mm-dd format, and where there is no month or day value, it adds "01"). In general, the variations in date format have no real impact on the use of dates to limit searches since the harvester software performs this post-harvest normalization. If a user limits her searches to a specific date range, the data in the CARL harvester's date field will be valid, although the user may be under the (possibly false, possibly true) impression that the date is the date of publication.

Type

Variances in the values of the uDC element "type" could potentially have a greater impact on users' queries than the variances observed in the use of the date element, since type is intended to indicate the genre of the work being described and may therefore be used to limit queries to specific kinds of documents. In the harvested metadata, the following values are present in Table VI.

It is obvious from this list that different terms are used to describe the same genre; for example, "Journal (Paginated)" "article" and "Journal (Online/Unpaginated)" all describe basically the same thing. French and English equivalents are also present, such as "Chapitre de livre" and "Book chapter".

Subject

Values occurring in "subject" elements include Library of Congress class numbers, subdivided Library of Congress Subject Headings, proper names of research centers and institutes, titles of conferences, textual dates (such as "April 25-27, 2002"), and what appear to be uncontrolled vocabulary keywords. Values appear in both English and French. Repeated keywords are sometimes given their own repeating elements and sometimes joined by commas or semicolons. Placeholders such as "No keywords supplied" are also supplied by at least one repository.

Description

Most "description" elements contain abstracts of the work being described. In some cases, French and English translations or English and Spanish translations of the same abstract are present. Another frequent type of values is short sentences such as "Commissioned by ..." or "Presented at ...". Phrases indicating that the work is a thesis, such as "Thesis (MA) – Faculty of Education" followed by the name of the university are also common. Occasionally, description elements also include the names of departments, institutes, or centers, or phrases that attribute some involvement with the creation of the work (such as "Video produced by ..."). In a few instances, the values of this element are "forthcoming" and "none."

Publisher

In some cases the "publisher" element contains the name of the university hosting the repository. In others, the name of the journal publisher is used, presumably if the item in the repository is an e-print. The names of institutes, centers, and other organizations that apparently had some role in the creation of the documents also appeared frequently.

Value	Frequency
Journal (paginated)	2,229
Text	886
Article	814
Electronic thesis or dissertation	202
Journal (on-line/unpaginated)	179
Thesis	172
Yes	134
Technical report	132
Nonpeerreviewed	86
Other	82
Video	61
Journal (online/unpaginated)	57
Book	52
Presentation	50
Working paper	48
Conference paper	42
Learning object	42
Book chapter	30
Dataset	17
Conference or Workshop Item	16
Image	16
Peerreviewed	13
No	9
Recording, oral	9
Présentation	8
Autre	7
Rapport technique/de consultation	7
Preprint	5
Monograph	3
Article de revue savante/scientifique	2
Chapitre de livre	2
Conference proceedings	2
Texte de conférence/séminaire	2
Article de quotidien/magazine	1
Livre	1
Objet d'apprentissage	1

Table VI.
Values used in the "type" element

Identifier

The "identifier" element is intended to identify the resource being described uniquely by means of a string or number conforming to a formal identification system, such as a URL, a Digital Object Identifier (DOI)[7], an ISBN, or an ISSN. In practically all cases, these are in fact the types of information used in identifier elements in the harvested metadata. The most common type is URL; cursory observation suggests that the destination of these URLs is either the record in the local repository or the document file itself. Handles[8] as assigned by DSpace are very common, occurring in about 76 percent of all instances of the identifier elements. DOIs are used very infrequently (only four occurrences intotal). Small numbers of identifier elements contain unformatted citations describing articles in print journals or annual reports, and some contain what appear to be locally-derived call or classification numbers

(but with no information associating them to a particular classification scheme, repository, or namespace).

Relation

"Relation" is commonly used in this data set to record a bibliographic citation describing the formally published version of the document, which is not a practice described in the Dublin Core documentation. One repository used this element to provide a link to the record for the work in its local library catalogue. Another repository included what appeared to be the complete reference list from the being described (i.e. brief citations to all the sources cited in the work).

The fact that the relation element is used at least to some extent in all but one repository's records suggests that repository implementers frequently map native metadata elements to it when formatting their native metadata schema as uDC for OAI harvesting, but the variances described in the previous paragraph suggests that the purpose of relation is interpreted inconsistently by repository implementers.

Other elements

Like the creator element, "contributor" is also used consistently by all repositories that include it (i.e. it generally contains people's names). "Format" is used is in accordance with best practice guidelines in the DCMI Metadata Terms document[9], that is to say, describing the format of the item using the Internet Assigned Numbers Authority MIME Media Types vocabulary, such as application/pdf, text/html, etc.[10] and, in repeated format elements, the size of the document's file in bytes. In general, the "language" element contains the RFC 3066[11] and ISO 639[12] two and three letter primary language codes, also recommended by the DCMI Metadata Terms guidelines, but one repository did not follow this practice and used the word "english" instead of the code "en". The "rights" element included brief, generic copyright statements such as "Copyright remains with the author" and "Copyright [author's name] 2003" and longer statements with a more legalistic tone (I hereby certify that, if appropriate, I have obtained and attached hereto a written permission . . .). For the "source" element, one site used the URL of the item in their repository, another used the name of a research unit on campus, and another used the name of what appeared to be a conference held on campus.

As mentioned at the beginning of this section, the use of the "creator" element is fairly standardized. However, this element can be used to illustrate one aspect of harvested metadata that becomes particularly important within a small group of repositories such as the one we are dealing with: the dominance of a repository platform can introduce a perceptible imbalance in the distribution of elements within the collection. For example, prior to the availability of DSpace Version 1.2, the use of DSpace Version 1.1 by a large number of repositories contributing to the CARL harvester resulted in a inflated number of Dublin Core "contributor" elements and an artificially low number of "creator" elements because that particular version of DSpace mapped its internal "author" field to contributor instead of creator, a practice recommended in the September 24, 2002 draft version of the Dublin Core Library Application Profile[13]. Subsequent versions of DSpace map their author field to creator in the OAI metadata, and as individual repositories have upgraded, the relative number of creator and contributor elements in the CARL harvester's database has come to more accurately reflect the metadata as it exists in the repositories' native metadata. Another example of DSpace's influence on the

aggregated metadata is that 76 percent of all identifier elements in this data set are handles, which DSpace automatically assigns.

Current activities

This snapshot of the metadata being collected by the CARL harvester reveals inconsistencies that bring into question its effectiveness as a resource discovery tool. CARL is investing in several strategies that address that question.

Development of an application profile

CARL is supporting a working group comprised of repository implementers from across Canada which is working to develop an application profile that can guide institutions contributing metadata to the CARL harvester. This group is active and plans to produce a full application profile by early 2006.

An application profile is a "set of metadata elements, policies, and guidelines defined for a particular application or implementation[14]." Within the context of the CARL harvester, such a profile will stipulate which elements are required, recommended, and optional, and will clarify best practices for the use of specific elements. Models that the CARL group could follow include the application profile developed by the OhioLink Digital Media Center[15], the "Western States Dublin Core Metadata Best Practices" document from the Colorado Digitization Program[16], and the application profiles listed in the DESIRE Registry based at UKOLN[17].

Some areas that will be considered for inclusion in an application profile for use at institutions participating in the CARL Project include:

- best practices in the use of elements that have a large impact on the success of subject queries, such as the description and subject elements;
- standardized language for use in the rights, source, and publisher fields; and
- standardized values for use in the type, identifier, rights, and language fields.

Vocabularies that could be used for this purpose may already exist, such as the DCMI Type Vocabulary[18], and other vocabularies may need to be created to suit the group's needs, particularly to account for the French and English metadata created by the contributing institutions. French-language and English-language institutions are contributing to the harvester, but the outcomes of research from any institution may appear in either language. It may prove challenging to use uDC to express metadata in both languages or to allow limiting searches to records in a particular record. In fact, it may be necessary to abandon use of uDC in favor of a richer metadata encoding schema if the expression of the language of an element's value is seen as necessary, since information about the language of an element's value can only be encoded in qualified Dublin Core using attributes such as $<$ subject lang $=$ "fr" $>$. Encoding this type of information in uDC can become very application-specific since doing so relies on formatting the element value in such a way that language indicators could be identified programmatically by harvester software or by the services built around the harvested metadata. However, it may be possible to take this approach using standardized techniques such as Dublin Core Structured Values (DCSV)[19].

One set of issues that makes developing and implementing application profiles for IRs challenging is those surrounding author-generated metadata. Libraries maintaining institutional or disciplinary repositories have to decide whether they

will apply quality control to author-generate metadata, and they have to balance the benefits of having authors create metadata and the difficulties of requiring or encouraging authors to adhere to standard practices or metadata creation tools that impose restraints on data input[20]. These issues are not specific to institutions contributing to the CARL harvester but many of them are supporting IRs in which authors submit their own documents and create at least initial versions of the associated metadata.

Application profiles should be developed so that they are as independent of particular harvesters and services as possible. Even though CARL is committed to using the PKP harvester as the basis for their search service, ideally the metadata being exposed by contributing repositories can be used effectively by a variety of different applications. The real challenge will be to develop an application profile that both accommodates current needs and also ensures that whatever effort we put into creating metadata are not wasted in the future. Also, any guidelines we devise for the local creation of metadata so it is optimized for centralized collection must respect local repositories' desire to create richer descriptions of documents than a given harvester may be able to collect, and acknowledge the need for some repositories to create and contribute very simple metadata that may not be as rich as other metadata in the harvester.

Improving the harvester application

As explained earlier, the PKP's Open Archives Harvester Software provides the platform for the CARL harvesting service. In its standard form, the PKP software is optimized for harvesting the metadata from the PKP's other software products, the OJS and the OCS. In order to improve the PKP harvester's utility for the CARL context, SFU Library staff removed the PKP harvester's "discipline" element from the search database and interface, and modified the source code that captured this element from OJS and OCS[21].

CARL has committed funds so that this change and others can be incorporated into the PKP harvester source code and contributed back to the project in accordance with the terms of the GNU General Public License[22], under which the PKP makes the harvester and its other products available. Work on this development will begin in June 2005 and will:

- Add options to allow harvester implementers to decide whether they wish to use the PKP-application specific metadata handling (for example, retain the "discipline" element in the harvester) or use a more generic uDC metadata schema, as described above.
- Increase the robustness of OAI harvesting. Currently the PKP harvester cannot harvest metadata from some repositories that use an HTTP "redirect" in their OAI gateway applications, and it does not support OAI-PMH flow control.
- Add options to allow implementers to configure search interface defaults and behaviors (currently, the default Boolean operator in the search interface is "or"); add to allow phrase searching.
- Add views of the harvested metadata that will allow browsing indexes of specific elements such as creator, title, subject, etc.

- Allow logging mechanisms that would facilitate analysis of how a harvester implementation is being used, such as search and browse logging, additions of new records, etc.
- Develop interfaces so other applications can interact with the harvester programmatically (for example, so other services can query the harvester via a standard GET URL).
- Integrate "newest items" list into home page.
- Integrate user login and personalization features, such as "saved searches" that will generate e-mails to the user when new records matching the saved search are added to the harvester.

Many of these changes have been suggested by users of the CARL harvester and by users of other implementations of the PKP harvester software. The administrators of CARL repositories have been particularly helpful in suggesting improvements, since they tend to pay close attention to how the metadata they are supplying works once it has been taken out of their local systems and placed within a new one.

Coincidentally, in January 2005, the Public Knowledge Project (based at the University of British Columbia), the SFU Library, and SFU's Canadian Centre for Studies in Publishing[23] entered into an agreement to move the support and continued development of all three of the PKP software applications to the SFU Library. Through its experience hosting the CARL harvester, the SFU Library is in an optimal position to improve the functionality of the already extremely useful PKP Open Archives Harvester.

Further investigation

Observation of the metadata harvested from the repositories participating in the CARL harvester raises some issues that are worth investigating more closely. The first issue surrounds the value of trying to improve the quality of the metadata at its source, the repositories. The focus so far in this paper has been on looking at patterns in the harvested metadata so we can identify areas that can be improved at the source of the metadata. This approach is not the only one that has been posited by researchers investigating ways to improve the usefulness of collections of harvested metadata. In his "Bitter Harvest: Problems & Suggested Solutions for OAI-PMH Data and Service Providers," Tennant (2004) proposes post-harvest normalization as a general strategy for overcoming some of the problems with metadata harvested via the OAI-PMH. As noted earlier, the PKP harvester software does perform some simple normalization on date elements, and investigating what other elements might be realistic candidates for this type of normalization within the group of repositories contributing to the CARL harvester might be a worthwhile exercise.

The apparent overlap observed between subject and description, publisher and description, source and identifier/description, and relation and publisher should be investigated more thoroughly, since doing so may assist us in developing best practices in assigning various types of content to specific elements. For example, by distinguishing between the names used in the publisher and relation elements in the harvested metadata, we might be able to define when it most appropriate to use one instead of the other, and if this distinction is important to end-users, then they will have more success in using search tools we build around the harvested metadata.

Another area that warrants further investigation, and that has only been alluded to so far, is query failure analysis. From the access logs of the web server hosting the CARL harvester, we know that between April 1, 2004 and June 15, 2005, users performed 1,628 searches (for an average of 3.7 searches per day). About 1,277 of those used the harvester's simple search interface and 351 used its advanced interface. Of all searches, 378 (app. 21 percent) returned no records. However, these few facts are all we know for certain about the queries that users are issuing. Inferring user behavior from the information currently tracked in the access logs of the web server hosting the CARL harvester would be rather speculative. Gathering reliable and useful information on how people are using the CARL harvester will be challenging because we have to define what aspects of user behavior we want to track before we implement the tracking mechanisms. However, doing so will allow us to tailor the application's functionality and the nature of the contributed metadata to better suit users' behavior.

Conclusion

Despite the common attitude that Google will allow researchers to find everything libraries put online easily and effectively, collecting metadata from related repositories and making it searchable in one interface allows us to create effective virtual collections and other services and to cater to the needs of specialized audiences. The value of the aggregated metadata in a centralized service such as the CARL harvester is only as high as the value of the metadata in the contributing repositories, and if libraries are going to convince researchers to use these tools instead of Google (or whatever general search engine replaces Google as the default web search engine), they have to invest in making the tools as effective as possible. It is only worth collecting metadata from a number of repositories and making it searchable, therefore, if we are committed to learning about the metadata and improving its quality.

The type of analysis used in this paper is one approach to discovering patterns in a set of metadata harvested for the purpose of allowing end-users to search it, and some of the patterns that have emerged may enable repository implementers to change the ways they create their metadata, or at least the ways they create the uDC representations of it that are collected by the CARL harvester.

Notes

1. http://carl-abrc-oai.lib.sfu.ca
2. www.openarchives.org
3. http://pkp.sfu.ca/pkp-harvester/
4. http://pkp.sfu.ca/ojs/
5. http://pkp.sfu.ca/ocs/
6. http://search.cpan.org/dist/OAI-Harvester/
7. www.doi.org
8. www.handle.net
9. www.dublincore.org/documents/dcmi-terms/
10. www.iana.org/assignments/media-types/
11. www.ietf.org/rfc/rfc3066.txt

12. www.loc.gov/standards/iso639-2/
13. The DSpace development team was following the best practice guideline in the (then current) working draft of the DC-Library Application Profile (http://dublincore.org/documents/2002/09/24/library-application-profile/). The latest version of the Profile clarifies this practice somewhat.
14. http://dublincore.org/documents/usageguide/glossary.shtml. For more information on application profiles, see Rachel Heery and Manjula Patel (2000), "Application profiles: mixing and matching metadata schemas" *Ariadne* No. 25, www.ariadne.ac.uk/issue25/app-profiles/
15. http://dmc.ohiolink.edu/docs/DMC_AP.pdf
16. www.cdpheritage.org/resource/metadata/wsdcmbp/index.html
17. http://desire.ukoln.ac.uk/registry/appprofile.php3
18. http://dublincore.org/documents/dcmi-type-vocabulary/
19. http://dublincore.org/documents/2000/07/11/dcmi-dcsv/#sec2
20. A useful discussion of the issues surrounding author-created metadata is Jane Greenberg *et al.*, " Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization" Journal of Digital Information, Vol. 2 No. 2 (2002), http://jodi.tamu.edu/Articles/v02/i02/Greenberg/.
21. The posting to the PKP Support Forum that details the changes made is available at http://pkp.sfu.ca/support/forum/viewtopic.php?t = 65
22. www.gnu.org/licenses/licenses.html#GPL
23. www.ccsp.sfu.ca

References

Moen, W.E. and Benardino, P. (2003), "Assessing metadata utilization: an analysis of MARC content designation use", paper presented at the 2003 Dublin Core Conference, Seattle, WA, September 28-October 2, 2003, available at: www.unt.edu/wmoen/publications/MARCPaper_Final2003.pdf

Tennant, R. (2004), "Bitter harvest: problems & suggested solutions for OAI-PMH data & service providers", www.cdlib.org/inside/projects/harvesting/bitter_harvest.html

Ward, J. (2004), "Unqualified Dublin Core usage in OAI-PMH data providers", *OCLC Systems & Services: International Digital Library Perspectives*, Vol. 20 No. 1, pp. 40-7.